



Article

Advancing DNA Language Models through Motif-Oriented Pre-Training with MoDNA

Weizhi An ¹, Yuzhi Guo ¹, Yatao Bian ², Hehuan Ma ¹, Jinyu Yang ¹, Chunyuan Li ¹ and Junzhou Huang ^{1,*}

¹ Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019, USA; weizhi.an@mavs.uta.edu (W.A.); yuzhi.guo@mavs.uta.edu (Y.G.); hehuan.ma@mavs.uta.edu (H.M.); jinyu.yang@mavs.uta.edu (J.Y.); chunyuan.li@mavs.uta.edu (C.L.)

² Tencent AI Lab, Shenzhen 518000, China; yataobian@tencent.com

* Correspondence: jzhuang@uta.edu

Abstract: Acquiring meaningful representations of gene expression is essential for the accurate prediction of downstream regulatory tasks, such as identifying promoters and transcription factor binding sites. However, the current dependency on supervised learning, constrained by the limited availability of labeled genomic data, impedes the ability to develop robust predictive models with broad generalization capabilities. In response, recent advancements have pivoted towards the application of self-supervised training for DNA sequence modeling, enabling the adaptation of pre-trained genomic representations to a variety of downstream tasks. Departing from the straightforward application of masked language learning techniques to DNA sequences, approaches such as MoDNA enrich genome language modeling with prior biological knowledge. In this study, we advance DNA language models by utilizing the Motif-oriented DNA (MoDNA) pre-training framework, which is established for self-supervised learning at the pre-training stage and is flexible enough for application across different downstream tasks. MoDNA distinguishes itself by efficiently learning semantic-level genomic representations from an extensive corpus of unlabeled genome data, offering a significant improvement in computational efficiency over previous approaches. The framework is pre-trained on a comprehensive human genome dataset and fine-tuned for targeted downstream tasks. Our enhanced analysis and evaluation in promoter prediction and transcription factor binding site prediction have further validated MoDNA's exceptional capabilities, emphasizing its contribution to advancements in genomic predictive modeling.

Keywords: transformer; self-supervised learning; genomic predictive modeling



Citation: An, W.; Guo, Y.; Bian, Y.; Ma, H.; Yang, J.; Li, C.; Huang, J.

Advancing DNA Language Models through Motif-Oriented Pre-Training with MoDNA. *BioMedInformatics* **2024**, *4*, 1556–1571. <https://doi.org/10.3390/biomedinformatics4020085>

Academic Editors: Tamir Tuller and Burghardt Wittig

Received: 27 February 2024

Revised: 10 May 2024

Accepted: 4 June 2024

Published: 12 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The exploration of non-coding regulatory genome regions, aiming to uncover the functionality of regulatory elements in gene expression coordination, has emerged as a focal point in recent research endeavors [1–3]. Remarkably, in humans, a mere fraction of DNA is dedicated to coding proteins, leaving the vast majority, approximately 98%, as non-coding DNA [4]. This non-coding segment harbors sequences pivotal to regulatory elements, which orchestrate the activation and suppression of genes across various contexts. These elements serve as docking sites for specialized proteins, thereby modulating gene expression [5]. The advent of deep learning computational methodologies has notably enhanced the predictive analysis of non-coding regulatory elements [6,7]. Intriguingly, researchers have identified linguistic-like statistical characteristics within non-coding DNA sequences, drawing parallels to natural languages [8]. This similarity between non-coding DNA sequences and human language has paved the way for the application of language models in decoding the cryptic language of non-coding DNA [9,10].

Various deep learning architectures, including Convolutional Neural Networks (CNNs) [11,12], have been employed to dissect gene regulatory mechanisms within genomic sequences [13,14]. Models such as Enformer [15], ImaGene [16], and DanQ [17] have

demonstrated notable success in differentiating promoter sequences, representing genomic information as abstract imagery, and predicting non-coding functions, respectively. Notably, Enformer [18] leverages Transformer-based models [19–22] to harness self-attention mechanisms, enabling the integration of extensive DNA contextual information and facilitating the interpretation of long-range dependencies within the genome.

Despite these advancements, significant challenges remain in effectively modeling and learning from DNA. First, the inherent limitation of CNNs in capturing long-range genomic dependencies poses a challenge. Second, the reliance on supervised learning constrains the models' generalization capabilities, tethering them to specific tasks defined by labeled data. This results in most sequence models being trained on a narrow scope of genome scenarios, which is susceptible to overfitting due to the scarcity of labeled datasets. Addressing these challenges necessitates an innovative approach to learning representations from unlabeled DNA sequences, capturing the intricate contextual information, and applying this knowledge across a spectrum of downstream tasks. This calls for a paradigm shift towards more adaptable and generalizable models, capable of transcending the limitations of current methodologies and unlocking new frontiers in genomic research.

Recent advancements by Ji et al. [23] have led to the adaptation of the Bidirectional Encoder Representations from Transformers (BERT) model [24,25] to genomic DNA settings, resulting in the development of DNABERT. This approach utilizes unlabeled human genome data to forge a versatile model capable of spanning a wide array of sequence-related tasks in a unified manner. While DNABERT harnesses the expansive genome data to capture a broad representation of DNA language, it may not fully integrate domain-specific knowledge, which is pivotal for a nuanced understanding of genomic functions. Directly applying BERT's methodology to DNA sequences—considering them analogous to human language—may overlook the unique structure and function inherent in genomic data. GeneBERT [26], another significant contribution to the field, introduces a multi-modal pre-training model pre-trained on millions of DNA sequences from the ATAC-seq dataset [27]. By treating different cell-type motif position weight matrices (PWMs) akin to image regions, GeneBERT attempts to blend sequence and structural data in its training process. However, this multi-modal approach may encounter limitations when applied to downstream tasks that solely involve sequence data, potentially restricting the model's applicability to purely sequence-based learning scenarios.

Our work builds on our previous MoDNA framework [28], presenting a more comprehensive evaluation. We integrate common DNA functional motifs as essential domain knowledge. The eukaryotic genome is replete with structured patterns such as repetitive elements, protein-binding sites, and splice sites [29], many of which manifest as motifs—short, recurring sequences indicative of specific protein-binding functions [30]. MoDNA leverages this concept by not only focusing on predicting masked k-mers but also incorporating a motif prediction task to enrich DNA representation learning. Through this dual-prediction approach in its self-supervised pre-training phase, MoDNA utilizes functional motifs to refine DNA representation embeddings, enhancing the model's biological relevance and predictive accuracy. In this study, we conduct a comprehensive analysis to demonstrate MoDNA's superior performance across various datasets, solidifying its effectiveness in genomic predictive modeling.

Our contributions with MoDNA are threefold: (1) By embedding domain knowledge into the learning process, MoDNA not only captures semantic representations of DNA but also accurately predicts motif occurrences, enriching the model with deep semantic insights from genomic sequences. (2) MoDNA marks the inaugural application of the ELECTRA [31] framework to DNA sequence representation, showcasing superior computational efficiency and performance metrics compared with BERT-based models. (3) Leveraging extensive unlabeled genome data, MoDNA demonstrates notable improvements in promoter prediction and transcription factor binding site identification tasks, setting new benchmarks against existing state-of-the-art models.

2. Related Work

2.1. Advancements in Self-Supervised Pre-Training for Natural Language Processing

The domain of Natural Language Processing (NLP) has been profoundly transformed by the advent of pre-trained language models, which adhere to a paradigm encompassing both pre-training and subsequent fine-tuning. Notable models such as GPT, BERT [24,32], RoBERTa [33], and ELECTRA [31] have demonstrated exemplary performance across a multitude of NLP tasks. These models employ self-supervised learning techniques on extensive corpora of unlabeled text, effectively circumventing the limitations inherent in supervised learning methodologies that necessitate laboriously annotated datasets. This innovative approach has been extended to the realm of biomedical language models, thereby facilitating the application of these advanced computational techniques to the analysis of genomic sequences [34,35].

Comparative Analysis of BERT and ELECTRA

The BERT model, in particular, has gained recognition for its capacity to leverage unlabeled data through the implementation of masked language learning alongside next-sentence prediction tasks, thereby enhancing the model's performance [36]. Nonetheless, the BERT framework is characterized by certain limitations, including a disproportionate focus on masked tokens to the detriment of non-masked tokens, a considerable computational demand owing to the model's reliance on a subset of input tokens for learning, and a notable discrepancy between the conditions prevalent in the pre-training and fine-tuning stages.

Conversely, the ELECTRA model proposes a novel solution to these challenges by adopting a generator–discriminator architecture, which optimizes the pre-training strategy. Within this framework, the generator fulfills a role akin to that of BERT by predicting masked tokens, whereas the discriminator evaluates the veracity of each token within the sequence, thereby enhancing the model's efficiency and reducing computational overhead. Furthermore, ELECTRA addresses the pre-training and fine-tuning discrepancy by ensuring consistency in the input tokens throughout both stages, thereby facilitating the model's adaptation to a diverse array of downstream tasks.

In the context of the present study, the ELECTRA model has been selected as the foundational framework, predicated on the hypothesis that DNA sequences can be effectively conceptualized and represented as sequential entities akin to linguistic texts. To the best of the authors' knowledge, this constitutes the inaugural application of the ELECTRA model to the domain of DNA sequence representation, thereby underscoring the model's versatility and potential applicability in the field of genomic analysis.

2.2. Tokenization of DNA Sequences

Despite the apparent parallels between DNA sequences and human language, the representation of DNA necessitates a bespoke approach. Traditional methods such as one-hot encoding, while straightforward, fail to encapsulate the complex semantic relationships inherent within DNA sequences. In contrast, the k-mer-based tokenization approach offers a more sophisticated alternative by segmenting sequences into overlapping substrings of length k. This methodology not only accommodates the canonical nucleotides—cytosine [C], guanine [G], adenine [A], and thymine [T]—but also incorporates special tokens ([CLS], [SEP], [PAD], [MASK], [UNK]) to fulfill various structural and functional roles within the modeling framework, thereby enriching the representational capacity of the model with respect to DNA sequences.

2.3. DNA Motifs

The intricate process of gene expression, particularly the transcription phase, is facilitated by the interaction of transcription factors with specific DNA sequences, known as enhancers and promoters. Within this biological context, DNA motifs—short, conserved sequence patterns endowed with significant biological functions—play a pivotal role [37,38]. These motifs, which may signify protein-binding sites or other functional elements, are

typically represented through Position Weight Matrices (PWMs) [39], providing a lucid and interpretable framework for their analysis. The MEME Suite [40], a comprehensive toolkit for motif-based sequence analysis, exemplifies the application of maximum likelihood algorithms for motif discovery in both DNA and protein sequences, thereby highlighting the synergistic relationship between computational modeling and biological research.

3. Methods

In this study, we introduce the MoDNA framework, which comprises two principal components: a generator and a discriminator. Both components are constructed using Transformer-based neural networks. This section elaborates on the architecture employed in our pre-training phase, the self-supervised tasks designed for model learning, and the subsequent fine-tuning process for genomic applications.

3.1. Model Architecture

The architecture of MoDNA is anchored by two networks: the generator and the discriminator. Each is structured with multiple Bidirectional Transformer Encoder blocks, as illustrated in Figure 1. We denote the number of Transformer blocks as L and the size of the hidden layers as H .

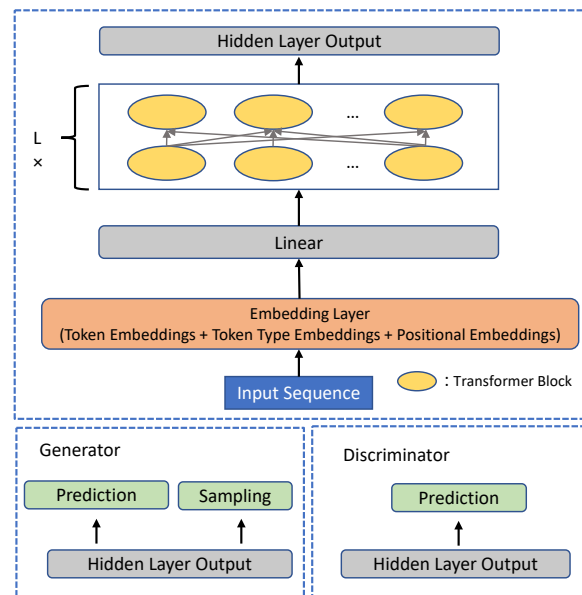


Figure 1. The structure of generator and discriminator.

A shared embedding layer forms the initial stage of both networks, incorporating token embeddings, token-type embeddings, and positional embeddings. This layer is pivotal in enabling the model to interpret the position and type of each token within a sequence. The Transformer blocks utilize a scaled dot-product attention mechanism, which is instrumental in directing the model's focus throughout the input data. The attention process is executed by transforming the input into query (Q), key (K), and value (V) matrices and computing the attention scores as follows:

$$\text{Att} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (1)$$

where d_k represents the dimensionality of the keys and queries. This mechanism allows the model to prioritize relevant parts of the input. The generator is tasked with masked genome modeling (MGM), operating in conjunction with the discriminator to refine its predictive capabilities.

3.2. Pre-Training Strategy

The pre-training of MoDNA involves a meticulously crafted two-step process, depicted in Figure 2. Initially, DNA sequence k-mers are randomly masked, employing a [MASK] token. The generator undertakes the challenge of predicting these masked segments. Concurrently, it generates new samples to substitute the masked tokens. The discriminator is then trained to discern whether each token in the sequence is original or has been replaced. This discrimination task is crucial for the model's learning, as it enhances the understanding of DNA sequence structures.

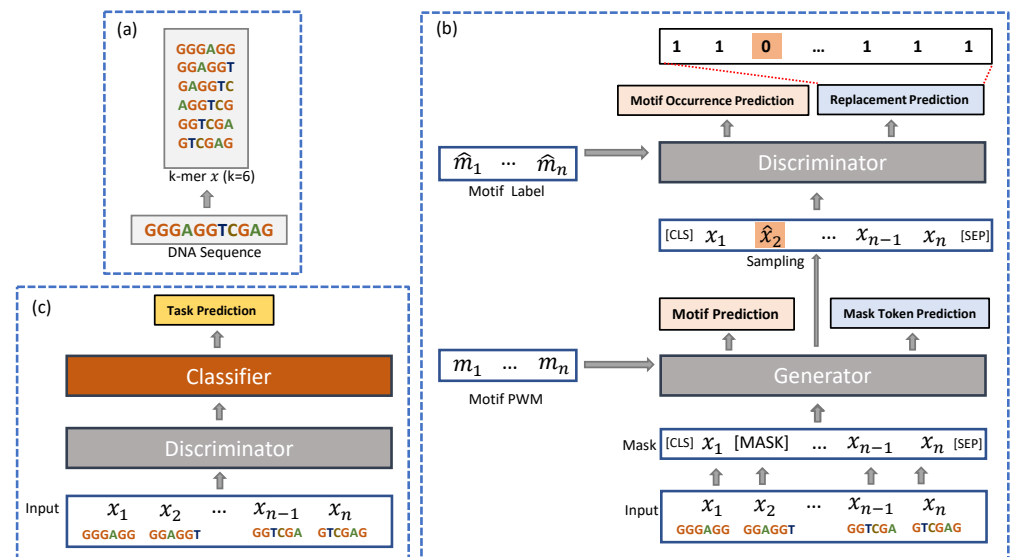


Figure 2. Overview of the MoDNA framework. (a) **DNA Sequence Representation:** Illustration of DNA sequence k-mers ($k = 6$), representing the basic units for analysis. (b) **Pre-training Pipeline of MoDNA:** The process begins with the random masking of input DNA sequence k-mers, with x_2 representing the masked token. DNA k-mer tokens, along with special tokens, are constructed into a sequence of DNA tokens. These tokens are input into the generator, which aims at two main objectives: predicting the masked genomic sequences and identifying motif patterns. The generator also produces a sampling $\hat{x}_2$ to substitute the masked token [MASK]. This modified sequence, combined with the unaltered tokens, is then processed by the discriminator, which is trained to detect replaced tokens and, with the given motif occurrence labels, to predict the presence of motifs. (c) **Fine-Tuning Pipeline of MoDNA:** The pre-trained discriminator's weights are used as the starting point. An additional multilayer perceptron is integrated for fine-tuning the model to specialize in various downstream tasks.

A significant aspect of our pre-training is the development of self-supervision tasks. Given the frequent occurrence of motifs—short, biologically significant patterns—in DNA sequences, we designed tasks that enable the model to recognize these motifs. We approach motif identification as a multi-label classification challenge, allowing the model to predict multiple motifs within a sequence simultaneously.

3.2.1. Genome Token Prediction

During the pre-training phase, our approach adopts the k-mer representation for unlabeled pre-training data, akin to the strategy employed by DNABERT. To illustrate, consider a DNA sequence *TTGGAAAT*; its six-mer representations would include sequences like *TTGGAA*, *TGGAAA*, and *GGAAAT*. The k-mer vocabulary encompasses all possible k-mer permutations, augmented by five special tokens. We transform the k-mer sequence $x = [x_1, \dots, x_n]$ into a tokenization embedding E , aligned with a comprehensive $k^4 + 4$ token vocabulary.

In alignment with the ELECTRA framework, our methodology entails the concurrent training of a generator and a discriminator. Prior to the integration of k-mer x into the tokenization embedding, we introduce randomness by masking six consecutive positions within the six-mer x , serving as the input for the generator. We denote these random mask positions as $c = [c_1, \dots, c_t]$, and accordingly, the tokens at these positions are substituted with [MASK] tokens, as demonstrated in the following equation:

$$r = \text{replace}(x, c, [\text{MASK}]). \quad (2)$$

The generator, upon receiving the masked input r , encodes it into a contextual embedding and undertakes masked genome modeling (MGM) to infer the original identities of the masked tokens. The MGM process is guided by the following loss function:

$$L_G = -\sum_{i \in c} \log p_G(x_i | r). \quad (3)$$

Subsequent to MGM, the replacements are sampled from the prediction output p to establish a sample distribution, as follows:

$$\hat{x} \sim p_G(\hat{x} | r). \quad (4)$$

The discriminator's inputs are then formulated by substituting the masked tokens x_i with the generated samples $\hat{x}$, as follows:

$$x^R = \text{replace}(x, c, \hat{x}). \quad (5)$$

The discriminator D plays a pivotal role in discerning the origin of each token x_i^R within the sequence x^R , determining whether it is derived from the original sequence or is synthetically generated by the generator. This discrimination task is quantified through the following loss functions:

$$L_D = \begin{cases} -\log D(x_i^R, x^R) & \text{if } x_i^R = x^R, \\ -\log (1 - D(x_i^R, x^R)) & \text{if } x_i^R \neq x^R. \end{cases} \quad (6)$$

This structured approach to genome token prediction underpins our pre-training strategy, facilitating a robust foundation for the subsequent fine-tuning on genomic tasks. Through the meticulous design of our generator and discriminator, we ensure that our model not only learns to accurately predict masked tokens but also gains a nuanced understanding of the genomic sequences, paving the way for significant advancements in genomic research.

3.2.2. Motif Prediction

Motifs, as recurrent nucleotide sequence patterns, hold significant biological implications, particularly in the context of gene regulation. The identification and interpretation of motifs can shed light on the biological functions associated with specific sequences. To extract motifs from DNA sequences [41], we utilize the MEME Suite [40], representing the discovered motifs as Position Weight Matrices (PWMs), denoted by $m = [m_1, \dots, m_n]$. These PWMs, corresponding to the unmasked portions of the input x , serve as critical indicators of the underlying biological functions.

Anticipating that our generator G can capture the distribution of these motifs, we incorporate PWMs m as labels for motif learning. This approach enables the generator to align its predictions with the biologically meaningful patterns encoded within the PWMs.

The motif prediction task hinges on the generator's ability to match the PWMs with the actual motif distribution within the sequences. This matching process is quantified

using the Kullback–Leibler Divergence (KL) loss [42], which assesses the similarity between two probability distributions. The motif learning loss, formulated as follows:

$$L_{G_{motif}} = \sum_{x \in c} m_i \log \frac{m_i}{W^T(h_G(x_i))} \quad (7)$$

relies on the generator’s last hidden layer representation $h_G(x)$ to predict the motif patterns.

Furthermore, the discriminator component, denoted as D_{motif} , is tasked with identifying motif occurrences within the sequences x^R , which incorporate the generated samples. The objective is for the discriminator to pinpoint the locations of motifs within the genome tokens, guided by the binary motif occurrence labels $\hat{m} \in (0, 1)$. The loss function for this task is as follows:

$$L_{D_{motif}} = - \sum_{i \in n} \hat{m}_i \log \text{sigmoid}(h_D(x_i)) + (1 - \hat{m}_i) \log (1 - \text{sigmoid}(h_D(x_i))). \quad (8)$$

3.2.3. Pre-Training Objectives

Our pre-training strategy encompasses the simultaneous training of the generator and discriminator, leveraging the motif prediction tasks as key self-supervised learning objectives. The composite loss function, integrating the losses from both components, is expressed as follows:

$$L = \lambda L_G + \alpha L_D + \beta (L_{G_{motif}} + L_{D_{motif}}). \quad (9)$$

Here, α and β are hyperparameters that calibrate the contributions of the respective loss components.

It is crucial to recognize the nuanced distinction in the operation of our discriminator D compared with traditional BERT models. Unlike BERT, which focuses solely on predicting masked tokens, our discriminator evaluates all tokens within the input sequences x^R , enriched by the generator’s samples. This comprehensive analysis, combined with the integration of motif patterns in both the generator and the discriminator, equips MoDNA with the capability to derive semantic and functional insights from the sequence data, thereby enhancing the overall quality of sequence representation learning.

3.3. Fine-Tuning

Following the pre-training phase, we enter the supervised fine-tuning stage, utilizing the discriminator network that has been pre-trained with the MoDNA framework. This step involves adapting our discriminator model, which has been primed on a vast array of genomic sequences, to specific downstream tasks that require precise and task-specific predictions.

A notable challenge encountered in traditional pre-training models like BERT is the discrepancy between pre-training and fine-tuning stages, primarily due to the use of masked inputs during pre-training and the transition to unmasked inputs for fine-tuning. MoDNA addresses this issue innovatively by employing generated samples to replace masked tokens during pre-training, ensuring that the discriminator’s inputs during fine-tuning encompass the complete set of input tokens without relying solely on masked portions. This approach allows for a more comprehensive training of the discriminator across all tokens, enhancing its predictive capabilities.

The utility of the MoDNA model extends to a variety of downstream tasks [43]. In this study, we focus on two critical applications: promoter prediction and transcription factor binding site (TFBS) prediction. For each task, we augment the pre-trained discriminator with a task-specific linear classifier, enabling the model to leverage the rich representations learned during pre-training for targeted predictions. During the fine-tuning process, we adjust all parameters of the pre-trained model in accordance with the specific requirements of each task, guided by task-relevant datasets and loss functions. This fine-tuning phase, spanning several epochs, culminates in the MoDNA model demonstrating enhanced per-

formance and predictive accuracy on these downstream tasks, showcasing its effectiveness in genomic sequence analysis.

4. Experimental Results

This section outlines the experimental framework of our study, detailing the procedures for both pre-training and fine-tuning phases, and introduces the datasets employed. Subsequently, we compare our results with leading methods in the field for various downstream tasks.

4.1. Pre-Training and Fine-Tuning Experimental Pipelines

For our pre-training data, we turned to the human genome DNA sequences from the GRCh38 assembly. To add a layer of biological context, motif scanning was performed on these sequences, resulting in the generation of Position Weight Matrix (PWM) representations for the motifs discovered. In a manner similar to DNABERT, our MoDNA model was pre-trained with a 15% masking rate for the six-mers within the sequences. This involved breaking down DNA sequences (with a maximum length of 512 nucleotides) into six-mer permutations. These six-mers were then randomly masked and supplied to the generator for processing. In tandem, PWMs and motif occurrence labels were integrated into the training regime of both the generator and the discriminator, enriching the learning process with biologically relevant information.

4.2. Implementation Details

In our pre-training, we use the Adam optimizer as the pre-training optimizer, and the batch size in the pre-training stage is 64. Like DNABERT, we warm up the model at the beginning of the first 10k steps. After that, the learning rate starts from 2×10^{-4} and linearly decays. The detailed parameters of our MoDNA are listed in Table 1. For our pre-training setup, the computations were carried out on four NVIDIA (NVIDIA Corp., 2001 Walsh Ave, Santa Clara, CA, USA) Quadro RTX 8000 GPUs, each with 48 GB of memory. The pre-training phase was completed in approximately 10 days given the substantial model size of 110 million parameters.

Table 1. Hyperparameters of pre-training on MoDNA.

Hyperparameter	MoDNA
Layers (L)	12
Hidden size (H)	256
Embedding size	128
Att-heads	4
Att-head size	64
Batch size	64
Attention dropout	0.1
Dropout	0.1
α	50
β	1

4.3. Datasets

4.3.1. Unsupervised Pre-Training Dataset

Our pre-training exploits the extensive human genome sequences from the GRCh38 assembly, sourced from the National Library of Medicine. Sequences containing ambiguous “N” nucleotides were excluded to ensure data quality, and the remaining sequences were adjusted to lengths within the 5–510 base pair range. This dataset includes a total of 4,963,161 sequences, providing a broad base for our model’s pre-training. For motif prediction, a set of 769 motifs, each 7–25 nucleotides long, was curated from HOCOMOCO v11 [44], a resource for human transcription factor binding motifs. These motifs were

then used to annotate the pre-training sequences, with the corresponding PWMs serving as labels.

4.3.2. Promoter Prediction Dataset

Our exploration into promoter prediction was conducted on a dataset derived from the Eukaryotic Promoter Database (EPDnew) [45], encompassing a total of 59,198 samples. Each sample, centered around a Transcription Start Site (TSS), consists of a 70-nucleotide sequence. For the fine-tuning phase, we allocated 10% of these samples for evaluation purposes, with the remainder utilized for training.

4.3.3. The 690 ChIP-seq Datasets

Our exploration further extended to the analysis of 690 ChIP-seq datasets, encompassing 161 transcription factors (TFs) across 91 human cell types [46,47]. The transcription factor binding site (TFBS) prediction task was conducted across all 690 datasets, employing 101 bp sequences centered on peak regions as positive samples. The generation of negative samples adhered to the methodology established by DESSO, selecting sequences with matching GC content from the GENCODE [48], thereby ensuring a balanced representation of genomic contexts.

4.4. Promoter Prediction

Promoter regions, situated proximal to Transcription Start Sites (TSSs), play a pivotal role in initiating the transcription process of DNA [6]. These regions often encompass critical short DNA elements and motifs, typically ranging from 5 to 15 bases in length, acting as binding sites for proteins that orchestrate the transcription initiation and regulation of downstream genes. For our promoter prediction task, we utilized the same core DNA promoter dataset as DNABERT, which consists of sequences that are 70 bp in length and centered around the TSS. In our comparative analysis, we benchmarked against the DNABERT pre-trained model, adhering to the fine-tuning strategy outlined in their study. Additionally, we drew comparisons with GeneBERT, a recent multi-modal pre-training model pre-trained on 17 million genomic sequences from the ATAC-seq dataset. Our experiments employed identical settings and datasets for fine-tuning MoDNA. Table 2 presents the performance metrics of GeneBERT, DNABERT, MoDNA without motif incorporation, and the complete MoDNA framework. We also compared our model with CNN, CNN+GRU, and CNN+LSTM. The results are presented in Figure 3. It is evident from the results that MoDNA achieves superior performance in promoter prediction tasks. Notably, despite GeneBERT's extensive training on large-scale genomic data and inclusion of cell-type-specific motifs, MoDNA surpasses its performance across all evaluation metrics, including Accuracy, AUC, F1, MCC, Precision, and Recall.

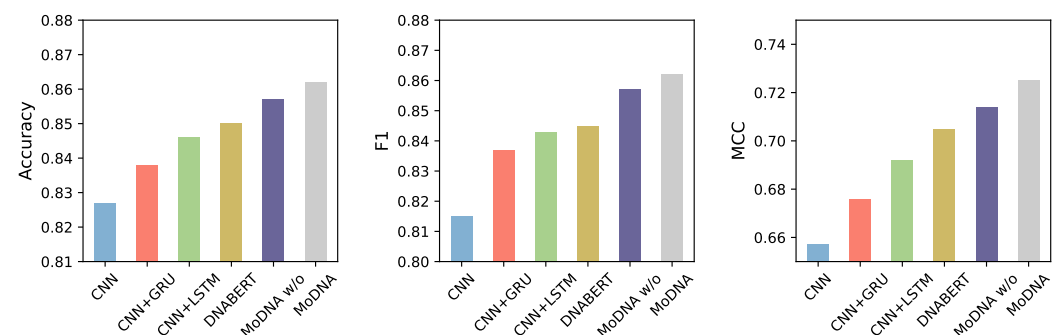


Figure 3. Comparison results on promoter core datasets.

Table 2. Comparison results on promoter prediction classification.

Method	Accuracy	AUC	F1	MCC	Precision	Recall
GeneBERT [26]	-	0.894	-	-	0.805	0.803
DNABERT [23]	0.841	0.925	0.840	0.685	0.844	0.841
MoDNA w/o motif	0.857	0.929	0.857	0.714	0.858	0.857
MoDNA (ours)	0.862	0.935	0.862	0.725	0.863	0.862

Furthermore, MoDNA demonstrates a noteworthy efficiency advantage over DNABERT, achieving a 1.5% relative improvement across all metrics. This enhancement underscores the effectiveness of MoDNA's pre-training strategy and the incorporation of self-defined motif prediction tasks. MoDNA's compact parameter set, in contrast with GeneBERT and DNABERT, contributes to its efficiency. Unlike traditional BERT pre-training approaches, MoDNA predicts across all tokens, thereby optimizing training efficiency and mitigating the data mismatch issue commonly encountered between pre-training and fine-tuning phases in previous models. Although our generator's pre-training includes masked genome prediction, we introduce a novel approach by generating mask tokens through sampling and incorporating them into the discriminator's input. This ensures consistency in input tokens throughout both pre-training and fine-tuning phases. Such an implementation not only showcases the efficacy of the ELECTRA framework but also validates our meticulously crafted pre-training tasks, which effectively harness implicit domain knowledge to bolster downstream task performance.

4.5. Transcription Factor Binding Site (TFBS) Prediction

Transcription involves copying a DNA segment into RNA, with transcription factor proteins playing a crucial role by binding to specific regulatory DNA regions. These proteins can attach to various regulatory elements such as promoters and enhancers, thus controlling gene expression. Given the pivotal role of these interactions in gene regulation, accurately predicting transcription factor binding sites is crucial for understanding DNA sequence functionality. To evaluate the efficacy of MoDNA in this domain, we fine-tuned our model on 690 ENCODE ChIP-seq datasets [47], as shown in Figure 4, comparing its performance against well-established methods in the field. Among these, DeepBind [1], a CNN-based model, stands out by learning sequence motifs as convolutional layer kernels, setting a high benchmark in TFBS prediction. Additionally, DeepSite [7] combines bidirectional long short-term memory with CNNs to capture the long-range dependencies between DNA sequence motifs. Another noteworthy method, DESSO [49], incorporates motif shape into DNA sequence representation learning with CNNs, demonstrating performance on par with DeepBind. Our comparative analysis spans all 690 datasets, with the average results encapsulated in Table 3. MoDNA distinguishes itself by outperforming the baseline methods across these datasets, achieving the highest average AUC and setting a new standard in TFBS prediction.

Table 3. Comparison results on transcription factor binding site classification.

Method	Accuracy	AUC	F1	MCC	Precision	Recall
DeepBind [1]	0.851	0.919	0.850	0.710	0.837	0.877
DeepSEA [50]	0.853	0.919	0.836	0.717	0.840	0.858
Basset [51]	0.741	0.860	0.685	0.531	0.799	0.729
DanQ [17]	0.840	0.910	0.823	0.694	0.848	0.823
DeepSite [7]	0.817	0.88	0.795	0.647	0.817	0.822
DESSO [49]	0.851	0.926	0.848	0.711	0.832	0.884
MoDNA w/o Pretrain	0.837	0.905	0.819	0.688	0.842	0.823
MoDNA (hltxtbfours)	0.856	0.935	0.851	0.727	0.859	0.856

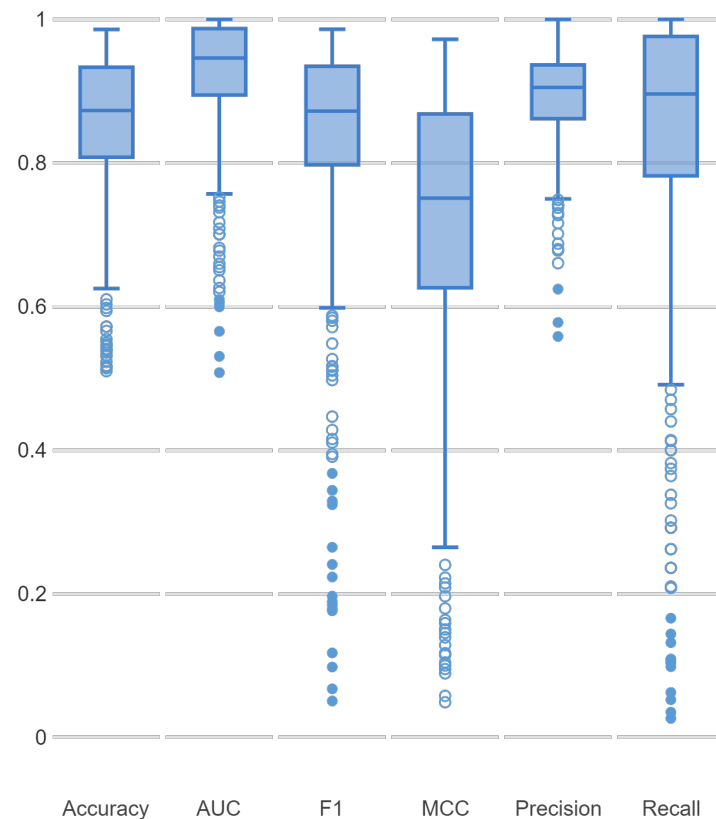


Figure 4. The performance of transcription factor binding sites of MoDNA in the 690 ChIP-seq datasets.

In addition to the comprehensive 690-dataset comparison, we conducted focused comparisons with DeepBind and GeneBERT for a more nuanced evaluation. DeepBind's performance was assessed on a subset of 506 ENCODE ChIP-seq datasets, where it demonstrated robustness by mitigating ChIP-seq data biases. GeneBERT, on the other hand, fine-tuned its model on nine CTCF site ChIP-seq profile datasets. Figures 5 and 6 present the comparative results, highlighting MoDNA's superior performance. On the 506 ENCODE ChIP-seq datasets, MoDNA's mean AUC reached 0.94, significantly outstripping DeepBind's 0.914. In the CTCF site experiment, MoDNA's mean AUC soared to 0.996, surpassing GeneBERT's 0.983 and underscoring the robustness of our pre-training and fine-tuning approach.

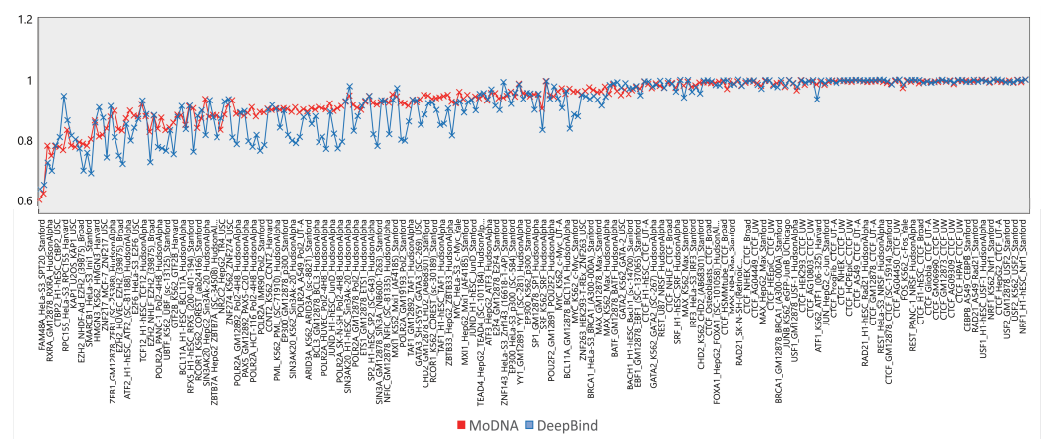


Figure 5. Comparison of AUC results with DeepBind of transcription factor binding site prediction on 506 TF binding profile datasets.

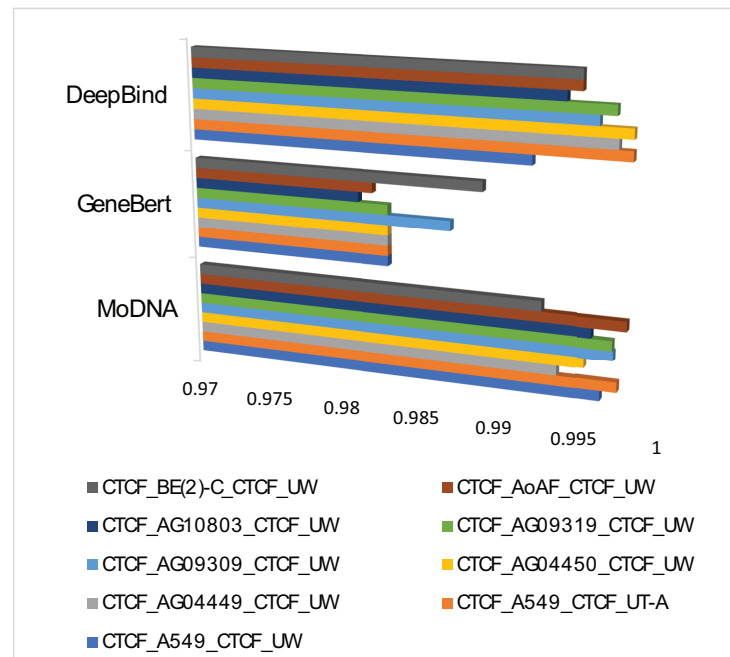


Figure 6. Comparison AUC results of transcription factor binding site classification on CTCF binding sites.

4.6. Ablation Studies and Discussion

To underscore the significance of our self-supervised pre-training approach, we conducted a comparative analysis of MoDNA's performance with and without the pre-training phase on the promoter prediction task. This comparison, detailed in Table 4, was conducted under identical experimental conditions. The results distinctly show a marked decrease in performance metrics for MoDNA without pre-training, underscoring the model's ability to internalize meaningful genomic representations during the pre-training phase. This phase enables MoDNA to assimilate essential biological insights, thereby enhancing its efficacy in downstream tasks. In a further exploration of MoDNA's efficiency, we pre-trained the model on a subset of the pre-trained dataset, which only pre-trains on 3000 DNA sequences, excluding motif prediction tasks. Remarkably, our MoDNA utilized a limited dataset and achieved an AUC of 0.926, surpassing GeneBERT's performance by 3.2%. A similar exercise with DNABERT yielded an AUC of 0.912, reinforcing MoDNA's superior model efficiency and its ability to achieve rapid convergence with limited data.

Table 4. Comparison between MoDNA with and without pre-training.

Method	Accuracy	AUC	F1	MCC	Precision	Recall
Pretrain	0.862	0.935	0.862	0.725	0.863	0.862
No Pretrain	0.808	0.889	0.808	0.618	0.809	0.809

To evaluate the impact of our motif prediction tasks, we also examined MoDNA's performance without the inclusion of motif-oriented learning. The experiments, utilizing the complete genome dataset, revealed that, while MoDNA without motif prediction outperformed GeneBERT and DNABERT, it fell short of the full MoDNA framework, as shown in Table 2. The enhanced performance of MoDNA, attributed to the incorporation of well-structured self-supervised tasks, validates our hypothesis that embedding motif prior knowledge into pre-training can significantly bolster the model's ability to learn functional biological features. Moreover, MoDNA's strategy of learning from all tokens—rather than just the masked ones as in BERT—proves to be advantageous. This approach, rooted in the ELECTRA model, not only improves training efficiency but also enriches the model with a deeper understanding of biological contexts through motif prediction tasks. This method-

ological shift is pivotal for adapting self-supervised learning to the unique challenges posed by biological data.

The prevailing paradigm in self-supervised learning, particularly within the NLP domain, hinges on reconstructing meaningful word pieces. However, the application of such NLP strategies to the “DNA language” is not straightforward due to the ambiguous nature of DNA sequences when treated as word pieces. This discrepancy raises a critical question: how can we leverage domain-specific prior knowledge to guide pre-training in fields distinct from NLP? MoDNA represents an initial foray into addressing this challenge, aiming to adapt self-supervised pre-training to a broader range of domain-specific issues. As we look to the future, the potential for incorporating additional functional feature tasks into MoDNA is vast. This expansion could further refine the model’s capabilities, paving the way for its application to a wider array of downstream tasks. The exploration of self-supervised pre-training, enriched with domain-specific knowledge, holds promise for advancing our understanding and application of machine learning in genomics and beyond.

5. Conclusions

In this work, we thoroughly evaluate the performance of MoDNA on diverse genomic tasks, demonstrating the framework’s superiority over existing DNA sequence language models that previously faced limitations due to their reliance on task-specific labeled data and constrained generalization capabilities. The quest for a versatile DNA language representation model remains a critical research endeavor. Our approach, MoDNA, extends beyond the straightforward application of NLP paradigms to DNA sequences by incorporating motif patterns during the pre-training stage. Unlike in NLP, where labeled data are more readily available, biological annotations often entail prohibitively expensive experimental procedures. MoDNA enhances computational efficiency for a given dataset size and resolves the data mismatch issue commonly associated with BERT by leveraging replacement tokens generated during pre-training. Furthermore, MoDNA’s discriminator is designed to process and differentiate among all input tokens, thereby optimizing learning efficiency.

In the realm of genomics, motifs are known to indicate protein binding sites within sequences. By embedding self-supervised motif prediction tasks that utilize motif patterns as domain knowledge, MoDNA enriches its pre-training phase. The generator focuses on motif prediction, while the discriminator assesses motif occurrences. This incorporation of biological insights enables MoDNA to capture a nuanced semantic representation of DNA. Upon fine-tuning, MoDNA demonstrates superior performance in promoter prediction and transcription factor binding site prediction, validating its effectiveness. The promising results invite the exploration of additional tasks, such as splice site prediction, to further challenge and refine MoDNA. Our findings underscore the value of integrating domain-specific knowledge through motif patterns into self-supervised learning tasks.

Although the MoDNA framework has demonstrated generalizability across different genomic datasets, its pre-training was conducted exclusively on the human reference genome. This approach limits the framework’s ability to account for sequence conservation and diversity across various species, potentially affecting its applicability to non-human genomic data. For future work, we aim to develop a more robust foundational model. This will involve extending our training datasets to include a diverse range of organisms and incorporating advanced algorithms that better capture genetic variability and conservation across species. Such enhancements will likely increase the predictive power and utility of the MoDNA framework in a broader array of genomic applications. Future research should also consider expanding the scope of self-supervised tasks to encompass other biologically relevant features, fostering a more comprehensive understanding of genomic sequences.

Author Contributions: Conceptualization, W.A. and Y.B.; methodology, W.A. and Y.G.; software, W.A. and Y.G.; validation, W.A.; formal analysis, W.A. and J.H.; investigation, W.A. and J.Y.; resources, J.H. and Y.G.; data curation, J.Y.; writing—original draft preparation, W.A. and Y.G.; writing—review and editing, H.M.; visualization, C.L.; supervision, J.H.; project administration, J.H.; funding acquisition, J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by US National Science Foundation IIS-2412195, CCF-2400785 and the Cancer Prevention and Research Institute of Texas (CPRIT) award (RP230363).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data derived from public domain resources. The datasets presented in this study are available in EPDnew at https://epd.epfl.ch/human/human_database.php?db=human, ENCODE database at <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/>.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Alipanahi, B.; Delong, A.; Weirauch, M.T.; Frey, B.J. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **2015**, *33*, 831–838. [CrossRef] [PubMed]
- Li, M.J.; Yan, B.; Sham, P.C.; Wang, J. Exploring the function of genetic variants in the non-coding genomic regions: Approaches for identifying human regulatory variants affecting gene expression. *Briefings Bioinform.* **2015**, *16*, 393–412. [CrossRef] [PubMed]
- Clauwaert, J.; Menschaert, G.; Waegeman, W. Explainability in transformer models for functional genomics. *Briefings Bioinform.* **2021**, *22*, bbab060. [CrossRef] [PubMed]
- The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **2007**, *447*, 799–816. [CrossRef]
- Andersson, R.; Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.* **2020**, *21*, 71–87. [CrossRef]
- Oubounyt, M.; Louadi, Z.; Tayara, H.; Chong, K.T. DeePromoter: Robust promoter predictor using deep learning. *Front. Genet.* **2019**, *10*, 286. [CrossRef]
- Zhang, Y.; Qiao, S.; Ji, S.; Li, Y. DeepSite: Bidirectional LSTM and CNN models for predicting DNA–protein binding. *Int. J. Mach. Learn. Cybern.* **2020**, *11*, 841–851. [CrossRef]
- Mantegna, R.N.; Buldyrev, S.V.; Goldberger, A.L.; Havlin, S.; Peng, C.K.; Simons, M.; Stanley, H.E. Linguistic features of noncoding DNA sequences. *Phys. Rev. Lett.* **1994**, *73*, 3169. [CrossRef]
- Brendel, V.; Busse, H. Genome structure described by formal languages. *Nucleic Acids Res.* **1984**, *12*, 2561–2568. [CrossRef]
- Corso, G.; Ying, Z.; Pándy, M.; Veličković, P.; Leskovec, J.; Liò, P. Neural Distance Embeddings for Biological Sequences. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 18539–18551.
- Liao, R.; Cao, C.; Garcia, E.B.; Yu, S.; Huang, Y. Pose-based temporal-spatial network (PTSNet) for gait recognition with carrying and clothing variations. In Proceedings of the Chinese Conference on Biometric Recognition, Shenzhen, China, 28–29 October 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 474–483.
- Guo, Y.; Wu, J.; Ma, H.; Huang, J. Self-supervised pre-training for protein embeddings using tertiary structures. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 6801–6809. [CrossRef]
- Yang, M.; Huang, H.; Huang, L.; Zhang, N.; Wu, J.; Yang, H.; Mu, F. Integrating convolution and self-attention improves language model of human genome for interpreting non-coding regions at base-resolution. *Nucleic Acids Res.* **2021**, *50*, e81. [CrossRef]
- Strodthoff, N.; Wagner, P.; Wenzel, M.; Samek, W. UDSMProt: Universal deep sequence models for protein classification. *Bioinformatics* **2020**, *36*, 2401–2409. [CrossRef] [PubMed]
- Umarov, R.; Kuwahara, H.; Li, Y.; Gao, X.; Solovyev, V. Promoter analysis and prediction in the human genome using sequence-based deep learning models. *Bioinformatics* **2019**, *35*, 2730–2737. [CrossRef]
- Torada, L.; Lorenzon, L.; Beddis, A.; Isildak, U.; Pattini, L.; Mathieson, S.; Fumagalli, M. ImaGene: A convolutional neural network to quantify natural selection from genomic data. *BMC Bioinform.* **2019**, *20*, 337. [CrossRef]
- Quang, D.; Xie, X. DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* **2016**, *44*, e107. [CrossRef] [PubMed]
- Avsec, Z.; Agarwal, V.; Visentin, D.; Ledsam, J.R.; Grabska-Barwinska, A.; Taylor, K.R.; Assael, Y.; Jumper, J.; Kohli, P.; Kelley, D.R. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **2021**, *18*, 1196–1203. [CrossRef] [PubMed]
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.

20. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. *arXiv* **2021**, arXiv:2103.00020.
21. Gao, S.; Alawad, M.; Young, M.T.; Gounley, J.; Schaefferkoetter, N.; Yoon, H.J.; Wu, X.C.; Durbin, E.B.; Doherty, J.; Stroup, A.; et al. Limitations of Transformers on Clinical Text Classification. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 3596–3607. [[CrossRef](#)]
22. Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; Wei, Y.; Huang, W.; Huang, J. Self-supervised graph transformer on large-scale molecular data. *arXiv* **2020**, arXiv:2007.02835.
23. Ji, Y.; Zhou, Z.; Liu, H.; Davuluri, R.V. DNABERT: Pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **2021**, *37*, 2112–2120. [[CrossRef](#)] [[PubMed](#)]
24. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
25. Min, S.; Park, S.; Kim, S.; Choi, H.S.; Lee, B.; Yoon, S. Pre-training of deep bidirectional protein sequence representations with structural information. *IEEE Access* **2021**, *9*, 123912–123926. [[CrossRef](#)]
26. Mo, S.; Fu, X.; Hong, C.; Chen, Y.; Zheng, Y.; Tang, X.; Shen, Z.; Xing, E.P.; Lan, Y. Multi-modal Self-supervised Pre-training for Regulatory Genome Across Cell Types. *arXiv* **2021**, arXiv:2110.05231.
27. Domcke, S.; Hill, A.J.; Daza, R.M.; Cao, J.; O'Day, D.R.; Pliner, H.A.; Aldinger, K.A.; Pokholok, D.; Zhang, F.; Milbank, J.H.; et al. A human cell atlas of fetal chromatin accessibility. *Science* **2020**, *370*, eaba7612. [[CrossRef](#)] [[PubMed](#)]
28. An, W.; Guo, Y.; Bian, Y.; Ma, H.; Yang, J.; Li, C.; Huang, J. MoDNA: Motif-oriented pre-training for DNA language model. In Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Northbrook, IL, USA, 7–10 August 2022; pp. 1–5.
29. Boeva, V. Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. *Front. Genet.* **2016**, *7*, 24. [[CrossRef](#)] [[PubMed](#)]
30. D'haeseleer, P. What are DNA sequence motifs? *Nat. Biotechnol.* **2006**, *24*, 423–425. [[CrossRef](#)] [[PubMed](#)]
31. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv* **2020**, arXiv:2003.10555.
32. Yamada, K.; Hamada, M. Prediction of RNA-protein interactions using a nucleotide language model. *Bioinform. Adv.* **2021**, *2*, vbac023. [[CrossRef](#)]
33. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
34. Wang, B.; Xie, Q.; Pei, J.; Chen, Z.; Tiwari, P.; Li, Z.; Fu, J. Pre-trained Language Models in Biomedical Domain: A Systematic Survey. *arXiv* **2021**, arXiv:2110.05006.
35. Choi, D.; Park, B.; Chae, H.; Lee, W.; Han, K. Predicting protein-binding regions in RNA using nucleotide profiles and compositions. *BMC Syst. Biol.* **2017**, *11*, 16. [[CrossRef](#)] [[PubMed](#)]
36. Wang, S.; Guo, Y.; Wang, Y.; Sun, H.; Huang, J. Smiles-bert: Large scale unsupervised pre-training for molecular property prediction. *BCB* **2021**, 429–436.
37. Bailey, T.L.; Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. *ISMB* **1994**, *2*, 28–36. [[PubMed](#)]
38. Yan, F.; Powell, D.R.; Curtis, D.J.; Wong, N.C. From reads to insight: A hitchhiker's guide to ATAC-seq data analysis. *Genome Biol.* **2020**, *21*, 22. [[CrossRef](#)] [[PubMed](#)]
39. Das, M.K.; Dai, H.K. A survey of DNA motif finding algorithms. *BMC Bioinform.* **2007**, *8*, S21. [[CrossRef](#)] [[PubMed](#)]
40. Bailey, T.L.; Boden, M.; Buske, F.A.; Frith, M.; Grant, C.E.; Clementi, L.; Ren, J.; Li, W.W.; Noble, W.S. MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res.* **2009**, *37*, W202–W208. [[CrossRef](#)] [[PubMed](#)]
41. Janky, R.; Verfaillie, A.; Imrichova, H.; Van de Sande, B.; Standaert, L.; Christiaens, V.; Hulselmans, G.; Herten, K.; Naval Sanchez, M.; Potier, D.; et al. iRegulon: From a gene list to a gene regulatory network using large motif and track collections. *PLoS Comput. Biol.* **2014**, *10*, e1003731. [[CrossRef](#)] [[PubMed](#)]
42. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [[CrossRef](#)]
43. Frazer, J.; Notin, P.; Dias, M.; Gomez, A.; Min, J.K.; Brock, K.; Gal, Y.; Marks, D.S. Disease variant prediction with deep generative models of evolutionary data. *Nature* **2021**, *599*, 91–95. [[CrossRef](#)] [[PubMed](#)]
44. Kulakovskiy, I.V.; Vorontsov, I.E.; Yevshin, I.S.; Sharipov, R.N.; Fedorova, A.D.; Rumynskiy, E.I.; Medvedeva, Y.A.; Magana-Mora, A.; Bajic, V.B.; Papatsenko, D.A.; et al. HOCOMOCO: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **2018**, *46*, D252–D259. [[CrossRef](#)]
45. Dreos, R.; Ambrosini, G.; Cavin Périer, R.; Bucher, P. EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic Acids Res.* **2013**, *41*, D157–D164. [[CrossRef](#)] [[PubMed](#)]
46. Lanchantin, J.; Sekhon, A.; Singh, R.; Qi, Y. Prototype Matching Networks for Large-Scale Multi-label Genomic Sequence Classification. *arXiv* **2017**, arXiv:1710.11238.
47. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489*, 57. [[CrossRef](#)] [[PubMed](#)]
48. Harrow, J.; Frankish, A.; Gonzalez, J.M.; Tapanari, E.; Diekhans, M.; Kokocinski, F.; Aken, B.L.; Barrell, D.; Zadissa, A.; Searle, S.; et al. GENCODE: The reference human genome annotation for the ENCODE Project. *Genome Res.* **2012**, *22*, 1760–1774. [[CrossRef](#)] [[PubMed](#)]

49. Yang, J.; Ma, A.; Hoppe, A.D.; Wang, C.; Li, Y.; Zhang, C.; Wang, Y.; Liu, B.; Ma, Q. Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework. *Nucleic Acids Res.* **2019**, *47*, 7809–7824. [[CrossRef](#)] [[PubMed](#)]
50. Zhou, J.; Troyanskaya, O.G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **2015**, *12*, 931–934. [[CrossRef](#)]
51. Kelley, D.R.; Snoek, J.; Rinn, J.L. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **2016**, *26*, 990–999. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.